Factorised Active Inference for Strategic Multi-Agent Interactions

Jaime Ruiz-Serra
Centre for Complex Systems,
The University of Sydney
Sydney, Australia
jaime.ruizserra@sydney.edu.au

Patrick Sweeney
Centre for Complex Systems,
The University of Sydney
Sydney, Australia
pswe2854@uni.sydney.edu.au

Michael S. Harré
Centre for Complex Systems,
The University of Sydney
Sydney, Australia
michael.harre@sydney.edu.au

Abstract

Understanding how individual agents make strategic decisions within collectives is important for advancing fields as diverse as economics, neuroscience, and multi-agent systems. Two complementary approaches can be integrated to this end. The *Active Inference framework* (AIF) describes how agents employ a generative model to adapt their beliefs about and behaviour within their environment. *Game theory* formalises strategic interactions between agents with potentially competing objectives. To bridge the gap between the two, we propose a factorisation of the generative model whereby each agent maintains explicit, individual-level beliefs about the internal states of other agents, and uses them for strategic planning in a joint context. We apply our model to iterated general-sum games with two and three players, and study the ensemble effects of *game transitions*, where the agents' preferences (game payoffs) change over time. This non-stationarity, beyond that caused by reciprocal adaptation, reflects a more naturalistic environment in which agents need to adapt to changing social contexts. Finally, we present a dynamical analysis of key AIF quantities: the variational free energy (VFE) and the expected free energy (EFE) from numerical simulation data. The ensemble-level EFE allows us to characterise the basins of attraction of games with multiple Nash Equilibria under different conditions, and we find that it is not necessarily minimised at the aggregate level. By integrating AIF and game theory, we can gain deeper insights into how intelligent collectives emerge, learn, and optimise their actions in dynamic environments, both cooperative and non-cooperative.

Keywords

free energy principle; game theory; theory of mind

ACM Reference Format:
Jaime Ruiz-Serra, Patrick Sweeney, and Michael S. Harré. 2025. Factorised Active Inference for Strategic Multi-Agent Interactions . In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

1 Introduction

Collective intelligence, the emergent ability of groups to solve problems more effectively than individuals, is a phenomenon observed across biological, social, and artificial systems. Understanding the mechanisms that drive this collective behaviour is essential for advancing fields as diverse as neuroscience, economics, and multi-agent systems.

Individual-level preferences incentivise the behaviour that shapes collective outcomes. While these preferences may not always conflict, tensions between cooperation and non-cooperation lead to emergent higher-level structures. Game theory models incentivised social interactions with potentially competing objectives, where a utility function maps behaviour to the real numbers. A Nash equilibrium represents the point where agents, independently maximising their utility, have no incentive to change their strategy.

Bridging the gap between idealised game-theoretic models and the often messy realities of agents interacting in complex environments presents a persistent challenge. Traditional game theory often falters when agents deviate from perfect rationality [10]. This challenge becomes particularly salient in the face of *strategic uncertainty*, where agents grapple with uncertainty about the actions and intentions of others, and *equilibrium selection*, where multiple potential equilibria exist without clear mechanisms for convergence. Shoham et al. [88] drew attention to a key issue with equilibrium selection: "It seems to us that sometimes there is a rush to investigate the convergence properties, motivated by the wish to anchor the central notion of game theory in some process, at the expense of motivating that process rigorously".

The *Active Inference framework* (AIF), a process theory rooted in neuroscience, can offer a compelling perspective on these challenges. AIF provides an empirically informed account of perception, action, and learning under uncertainty that has rapidly matured in recent years [24], but lacks a clear framework for multi-agent strategic interactions [17, 25]. Game theory often assumes perfect rationality and complete information, which AIF relaxes. Employing game theory as a model of incentivised decision-making, and AIF as the cognitive process underlying individual decisions allows for experiments with dynamic agent preferences while providing access to how their beliefs and precision change in response to others. Our results shed light on the mutual influence between individual cognition and structural dynamics at the collective level.

We begin by reviewing recent work on the intersection of AIF and Bayesian agents, game theory, and multi-agent systems (§2). Integrating the two ends of the spectrum, we propose a factorisation of the generative model whereby an agent maintains explicit, individual-level beliefs about the internal states of other agents and uses them for strategic planning in a joint context (§3).

We apply our model to iterated general-sum games with two and three players and study the ensemble effects of *game transitions*, where the agents' preferences (game payoffs) and their associated equilibria change over time (§4). We present a dynamical analysis of two key AIF quantities: the *variational free energy* (VFE) (§4.1) and the *expected free energy* (EFE) (§4.2) from numerical simulation

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

data. The ensemble-level EFE allows us to characterise the basins of attraction of games with multiple Nash Equilibria (such as the Stag Hunt) under different conditions, and we find that it is not necessarily minimised at the aggregate level [47].

2 Background

2.1 Iterated normal-form games

Iterated normal-form games (INFG) [41] provide a structured framework to study strategic interactions between agents, allowing for the analysis of decision-making processes over repeated encounters. INFG extend the basic framework of normal-form games, where players (agents) simultaneously choose strategies (actions), and their payoffs depend on the combination of all chosen strategies. In an iterated setting, this process repeats over multiple rounds, allowing agents to observe outcomes and potentially adapt their strategies over time. INFG are defined by a set of agents, a set of allowable actions for each agent, $\mathcal{U}_i$, a game payoff function, g, mapping every joint outcome (actions of all players involved) to a real number (payoff value for a given outcome)—thus encoding (often as a matrix) the incentives or preferences of the agents—, and possibly the total number of rounds (time steps) the game is played for. We use the term 'ego' to refer to any arbitrary agent from whose perspective we are describing the game, and 'alter' to refer to any other agent participating in the interaction.

The simplest games are (symmetric) two-player, two-action (2×2) games[1]. Here actions are $u \in \mathcal{U} = \{0, 1\} \equiv \{\mathsf{c}, \mathsf{d}\}$ ('cooperate' and 'defect') for each agent. Ego's payoffs for each of the four possible outcomes in these games are commonly referred to as *reward* (R) when both agents cooperate, *temptation* (T) when ego defects and alter cooperates, *sucker* (S) when ego cooperates but alter defects, and *penalty* (P) when both defect. From ego's point of view, her payoffs are:

$$\mathsf{g} = \begin{bmatrix} R & S \\ T & P \end{bmatrix} \tag{1}$$

Canonical games can be determined by the relative ordering of these payoff values [42], for example the well-known *Prisoner's Dilemma* (PD) has $T > R > P > S$, the *Chicken* game (Ch) has $T > R > S > P$, and the *Stag Hunt* (SH) has $R > T > P > S$. By setting each as an integer between 1 and 4 as in [17], we have:

$$\mathsf{PD} = \begin{bmatrix} 3 & 1 \\ 4 & 2 \end{bmatrix}, \ \mathsf{Ch} = \begin{bmatrix} 2 & 3 \\ 4 & 1 \end{bmatrix}, \ \mathsf{SH} = \begin{bmatrix} 4 & 1 \\ 3 & 2 \end{bmatrix}, \tag{2}$$

from which we can obtain the payoff function, e.g.

$$\mathsf{g}_{\mathsf{Ch}}(\mathsf{d}, \mathsf{c}) = T_{\mathsf{Ch}} = 4.$$

In the single-shot version of a normal-form game, agents typically maximise their payoff for that round alone. However, in iterated games, agents must consider long-term outcomes [31]. This opens up new possibilities for strategic behaviour, including learning (agents can learn from previous interactions adjusting their strategy to improve future outcomes), reciprocity (agents might cooperate if they believe others will reciprocate in future rounds, balancing short-term losses for long-term gains), and reputation and trust (the interaction history can influence future decisions, where agents may choose to trust or punish based on previous behaviour) [4].

[1]Two-action settings (e.g., cooperate/defect) are standard in game theory due to their simplicity, analytical tractability, and clarity of incentives. These settings focus on fundamental dynamics and are widely applicable to real-world scenarios abstracted into binary decisions.

2.2 Bayesian learning in games

Bayesian learning in strategic games builds on Savage's foundational work in Bayesian decision theory, which originally addressed games against nature [86]. In multi-agent settings, outcomes depend on the strategic interplay of agents, where each agent's actions influence and respond to those of others. This interdependence creates a dynamic, non-stationary environment, requiring agents to adapt continually as strategies coevolve [3, 45, 61].

The simulation literature focuses on how agents can learn equilibria through repeated interactions. The choice of priors significantly influences which equilibria agents achieve [16, 27, 68]. Under certain conditions, rational learning may converge asymptotically to a Nash equilibrium if agents' priors contain a 'grain of truth' [51, 52, 66]. A fundamental model in this area is fictitious play, where agents estimate opponents' strategies by averaging past actions and choosing their best response [9, 82]. This framework can be interpreted as sequential Bayesian inference, where each agent assumes opponents follow an unknown, independent, and stationary strategy [95]. Extensions of fictitious play introduce stochastic action selection [28, 30, 63], exponential forgetting [29], non-stationary strategies [87], and variational inference [80].

In AI, the multi-agent systems literature explores Bayesian methods for coordination and learning, though they have not yet gained the same traction as in single-agent reinforcement learning [32]. A common approach is to adapt single-agent algorithms, such as the Bayes-Adaptive Markov Decision Process [19] and its extensions [12, 39, 81, 83]. These algorithms, however, often struggle in multi-agent environments due to the non-stationarity introduced by agents' coevolving strategies, making it challenging for any single agent to converge to an optimal policy.

To address these challenges, type-based reasoning has emerged as a prominent approach for explicitly modelling other agents. In this approach, agents model others' behaviours as mappings from interaction histories to action probabilities [1, 44], allowing them to anticipate and respond to a wide range of strategies. This method addresses the heterogeneity of multi-agent systems by classifying agents according to their learning capabilities and information structures [61]. By starting with a prior over these types, agents systematically update their beliefs based on observed actions, refining their predictions and strategies as they gather more information about their opponents' behaviors [2, 11, 46, 90, 92].

Recursive reasoning builds on type-based methods by incorporating not only an agent's beliefs about others but also their beliefs about the beliefs of others, forming a hierarchical structure. This approach underlies models like the Interactive Partially Observable Markov Decision Process [34–36], where agents maintain and update beliefs about others' beliefs and strategies across multiple levels. By modelling nested beliefs, agents better anticipate others' actions, laying the foundation for Theory of Mind models that support more sophisticated and adaptive interactions [6, 18, 67, 76, 92].

Another line of research extends graphical models to multi-agent contexts, leveraging conditional independencies for efficient representation and inference. Multi-agent influence diagrams and graphical games capture dependencies among agents, enabling efficient computation by focusing on local interactions [54, 56, 57]. Similarly, action-graph games and expected utility networks optimise inference by structuring interactions around shared actions within agent subsets, which is particularly advantageous in sparsely coupled games [49, 50, 58].

Finally, the intersection of Bayesian learning and bounded rationality frames decision-making as constrained optimisation under uncertainty. Product Distribution Theory applies the Maximum Entropy principle [48] to derive equilibria where agents balance utility maximisation with computational costs [91]. Grünwald and Dawid demonstrate that maximising entropy and minimising worst-case expected loss are dual problems, structured as zero-sum games between a decision-maker and nature [38]. Thermodynamic Decision Theory expands these principles, integrating utility (energy) and information-processing costs (entropy) within a variational framework where agents minimise free energy. This approach naturally extends to variational Bayesian inference, enabling efficient, approximate posterior updates under bounded rationality constraints [74]. This framework generalises to risk-sensitive control [20, 53, 60, 69, 70] and adversarial contexts [71, 72], illustrating its versatility across diverse decision-making scenarios.

2.3 Game theory and Active Inference

This section outlines how AIF and game theory have been combined to model strategic decision-making in social interactions. Yoshida et al. [93] examine how individuals infer the intentions of others in the spatial Stag Hunt game, highlighting that people engage in recursive thinking about others' beliefs. This approach connects strategic thinking in game theory with Bayesian inference and bounded rationality in cognitive psychology, providing insights into how people make decisions in uncertain social environments.

Moutoussis et al. [65] develop a formal model of interpersonal inference based on active inference principles, where agents infer their partner's likely type (cooperative or defecting) by observing past actions and updating their beliefs. This demonstrates how the computational models used to describe individual decision-making can be extended to the complexities of social interaction, where understanding others' mental states is crucial.

Demekas et al. [17] build on this by showing how AIF agents can learn effective strategies in the iterated Prisoner's Dilemma by continuously updating beliefs about transition probabilities between game states. Their generative model tracks how learning rates influence strategy development, offering analytical clarity through belief updates. Hyland et al. [47] introduce the 'Free-Energy Equilibria' framework, extending the EFE to strategic contexts by conditioning predictions on the joint policies of agents. This framework merges Nash equilibria with bounded rationality, proposing that cooperation could emerge as agents align actions through joint free-energy minimization.

Fields and Glazebrook [21] further explore how physical interactions can be framed as games using the Free Energy Principle, drawing attention to the undecidability of achieving Nash equilibria in classical and quantum contexts. This complexity helps explain why real-world systems often fail to converge to stable outcomes.

Grounding strategic decision-making in AIF provides a more realistic model for social interactions, moving beyond the assumptions of perfect rationality and complete information in traditional game theory, with a foundation in neuroscience.

3 Model description

In INFG, N agents interact by selecting an action each time step with the goal of maximising their payoffs as determined by the game payoff function, g. The agents observe the actions taken by each of the agents (including themselves) in the preceding step in order to decide how to act in the current step.

In our model[2], these observations are perceived via different *modalities*, m, one for each agent. For example, for $N = 3$ agents $m \in \{i, j, k\}$, $|\mathcal{U}| = 2$ actions, and taking an egocentric perspective, each observation is $\mathbf{o} = (o_i, o_j, o_k) \in \{\mathsf{c}, \mathsf{d}\}^N$, with the first modality being *ego*'s action $o_{i,t} = u_{i,(t-1)}$, and subsequent modalities $o_{j,t} = u_{j,(t-1)}$ each pertaining to an *alter* (j, k). Further details are provided in §3.2.1. In the following, we omit time and agent indices where implied by context.

To act effectively in response to her counterparts, ego must take into account each of her opponents' propensity for playing each action at a given time (e.g. for j to play 'cooperate', or $p(u_j = \mathsf{c})$). This propensity is driven by the opponent's 'internal world', ψ_j, which is not observable to ego [62].

An appropriate way to model ψ_j is as a hidden state using a *Partially-Observable Markov Decision Process* (POMDP), where, each time step, agents infer the current hidden state $s \in \mathcal{S}$ based on an observation $o \in \mathcal{O}$, and can influence it through their actions $u \in \mathcal{U}$ to maximise their payoff. AIF distinguishes between the external (ontological) *generative process*, which represents the actual dynamics of the environment, and the internal (epistemic) *generative model* of each agent, which encodes its beliefs about those dynamics, and as such is a good match for the POMDP formalism [15, 89].

3.1 Generative Model

An agent's generative model consists of a joint distribution over hidden states, observations, policies (action sequences), and model parameters. The short timescale dynamics are encoded in

$$p(s_{0:t}, o_{0:t}|u_{0:t-1}) = p(s_0)\prod_{\tau=1}^{t} p(o_\tau|s_\tau)p(s_\tau|s_{\tau-1}, u_{\tau-1}),$$

including the agent's prior beliefs about the initial state of the world $p(s_0)$ encoded in $\mathbf{D}$; the transition model $p(s_{t+1}|s_t, u_t)$, encoded in $\mathbf{B}$; and the observation likelihood $p(o|s)$, encoded in $\mathbf{A}$. For discrete POMDPs, the distributions can be obtained from their encoding as a categorical distribution, e.g. $p(o|s) = \mathrm{Cat}(\mathbf{A})$ where $\mathbf{A}$ is a $|\mathcal{O}| \times |\mathcal{S}|$ matrix, $\mathbf{B}$ is a $|\mathcal{S}| \times |\mathcal{S}| \times |\mathcal{U}|$ tensor, and $\mathbf{D}$ a vector in the simplex $\Delta_{\mathcal{S}}$.

In our application to strategic interactions, ego infers the hidden state of each agent [6, 84] in a corresponding *factor*, $n \in \{i, j, k\}$, of the generative model. Ego must make vast simplifications in modelling each alter's true and complex internal world ψ_j—which

[2] Code available at GitHub/RuizSerra/factorised-MA-AIF

encompasses all the components in alter's generative model, encoding his *beliefs* ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{D}$, q; see §3.2), *preferences* ($\mathbf{C}$; see §3.3), and *constraints* (such as habits $\mathbf{E}$, or level of rationality β_1; see §3.4) [33], all depicted in Figure 1. Our approach is to model this internal world (external to ego) as a categorical distribution over different types, with parameters $\mathbf{s}$, placing the opponent type on the simplex $\Delta_{\mathcal{S}}$. In the simplest such model, which we adopt here, these types may be 'cooperator' or 'defector', and the opponent could be anywhere between the two (i.e., we model the opponent's propensity for playing each action)[3].

Ego further assumes that her opponents play the actions they mean to play, ruling out the possibility of 'trembling hand' imperfections. This simplifies our model such that there is no 'ambiguity' in the environment, so the likelihood model for each factor $\mathbf{A}_n$ is an identity matrix, or equivalently that the observation likelihood is a Kronecker delta distribution, $p(o_m|s_n) = \delta_{o_m,s_n}, \forall\, m = n \in \{i, j, k\}$.

3.2 Variational inference

Every time step, having observed the actions of each agent, $o_n = u_n$, ego has to infer the underlying $s_n \equiv p(u_n)$, for each factor. This entails updating her posterior beliefs $p_i(s_j|o_j)$ about each hidden state factor through Bayesian inversion. The true posterior may be intractable, so it is approximated by a variational posterior $q_i(s_j)$. This is achieved by minimising the *Variational Free Energy* (VFE), which is expressed as (having omitted the agent i and factor j indices for brevity):

$$F[q, o] = \underbrace{\mathbb{E}_{q(s)}[-\log p(o, s)]}_{\text{energy}} - \underbrace{H[q(s)]}_{\text{entropy}} \tag{3a}$$

$$= \underbrace{\mathrm{D}_{KL}\left[q(s) \,\middle\|\, p(s|o)\right]}_{\text{divergence}} - \underbrace{\log p(o)}_{\text{evidence}} \geq \underbrace{-\log p(o)}_{\text{evidence}}, \tag{3b}$$

with $p(s)$ the prior over hidden states, and $q(s)$ the variational posterior, or the agent's beliefs about hidden states. We model beliefs via a Dirichlet($\boldsymbol{\theta}$) distribution with variational parameters $\boldsymbol{\theta}$. We approximate the VFE using the following Monte Carlo sampling procedure: let $q_l \sim \text{Dirichlet}(\boldsymbol{\theta})$ be the l-th sample from the Dirichlet distribution. With L samples $\{q_l\}_{l=1}^{L}$, we can write an unbiased estimate of the VFE[4] as

$$F[q, o] = -\mathbb{E}_{q(s)}\left[\log p(o|s) + \log p(s) - \log q(s)\right] \tag{4a}$$

$$\approx \hat{F}\left[\{q_l\}_{l=1}^{L}, o\right] = \frac{1}{L}\sum_{l=1}^{L} F[q_l, o] \tag{4b}$$

and optimise the variational parameters $\boldsymbol{\theta}$ through stochastic gradient descent on $\hat{F}$. Having found $\boldsymbol{\theta}^*$ upon completing the procedure, we can recover the inferred posterior $q(s)$ as the expected value of Dirichlet($\boldsymbol{\theta}^*$), which serves as the optimal point estimate under a quadratic loss function [8].

Since variational inference occurs every time step, the prior $p(s)$ is obtained from the previously inferred state and the transition model $\mathbf{B}_u$, i.e.

$$p(s_t) \approx q(s_t|u_{t-1}) = \sum_{s_{t-1}} p(s_t|s_{t-1}, u_{t-1}) q(s_{t-1}) \tag{5}$$

3.2.1 Factorised model Generative models in previous AIF-adjacent applications to game theory [17, 22, 65] assume the state space is joint across all agents in the game (i.e. {cc, cd, ... dd} for their two agents). Is a mean-field factorisation of the variational posterior, $q(s_i)q(s_j)q(s_k)$, adequate in the game-theoretic context, or should a joint distribution, $q(s_i, s_j, s_k)$, be assumed [80]? That is, can the hidden states be considered independent of each other? Recall that, when our agents perform inference, they approximate the posterior $p(s|o) \approx q(s)$. In the case of repeated normal-form games, the hidden state s_j is (the parameterisation of) the posterior distribution over actions (policy) $q(u_j)$ of opponent j, and the observation o_j is the action u_j taken by this opponent. So inferring a distribution over a single opponent's hidden state involves finding a $q(s_j) \approx p(s_j|o_i, o_j, o_k)$.

The *Markov Blanket* (MB) of a node or set of nodes $\mathcal{J}$ is defined as the set of $\mathcal{J}$'s parents, $\mathcal{J}$'s children, and any other parents of $\mathcal{J}$'s children [78]. Under this definition, if we let $\mathcal{J}$ be the nodes 'inside' agent j, including $s_j = q(u_j)$, the MB consists of $\{o_i, o_j, o_k, u_j\}$, which 'shield' the internal states of j from the external world. In INFG, taking dynamics into account, this MB set is actually $\{u_{i(t-1)}, u_{j(t-1)}, u_{k(t-1)}, u_{jt}\}$. Thus, we can say that

$$s_j \perp \{s_i, s_k\} \;\big|\; \{u_{i(t-1)}, u_{j(t-1)}, u_{k(t-1)}, u_{jt}\},$$

i.e., s_j is *conditionally independent* of s_i and s_k (and any other node outside j) *given* $\{u_{i(t-1)}, u_{j(t-1)}, u_{k(t-1)}, u_{jt}\}$. Furthermore, at time t when i infers s_j, the action u_{jt} has not happened yet, so that node does not exist in the Dynamic Bayesian Network. Therefore, $q(s_j) \approx p(s_j|o_i, o_j, o_k)$, i.e. given the MBs of agents, their internal states are conditionally independent.

Accordingly, in our model, each agent i infers the hidden state for every factor $n \in \{i, j, k\}$ individually, and thus retains a collection of parameters $\boldsymbol{\theta} \in \mathbb{R}_+^{N \times |\mathcal{U}|}$, with N the number of factors (agents being tracked) and $|\mathcal{U}|$ the number of actions.

3.3 Preferences and planning

AIF extends Bayesian learning by integrating action and decision-making, allowing agents to actively reduce uncertainty and achieve goal-directed behaviour in dynamic environments. Agents plan and select actions that both gather information and satisfy their preferences over observations $p^*(o)$, encoded in $\mathbf{C}$ (such that $p^*(o) = \text{Cat}(\mathbf{C})$ in discrete settings). This requires counterfactual thinking: 'what would I be likely to observe if I were to do u_i?'. The generative model employed in inference can be used for planning by predicting future states (via the transition model $\mathbf{B}_{\hat{u}}$) and observations (via the likelihood $\mathbf{A}$) given counterfactual actions $\hat{u}_i$ (distinguished from actual actions u_i). In the following, we denote predictive variables with a bar, e.g. $\bar{o}_j$ is what i might observe j doing in the next time step.

Agents achieve the desired exploration-exploitation trade-off by selecting actions that minimise an *Expected Free Energy* (EFE), comprising *salience* and *pragmatic value* terms. In what follows, we describe these terms in more detail and adapt them to INFG.

[3] The support of this distribution need not be limited to the number of actions. For example, an agent could consider four possible types for policy length 2, or more abstract types such as 'Tit for Tat' [4].

[4] The chosen form of the VFE for optimisation is derived from (3a) by replacing $p(o, s) = p(o|s)p(s)$ and subsuming all terms into a single expectation operator in (4a).

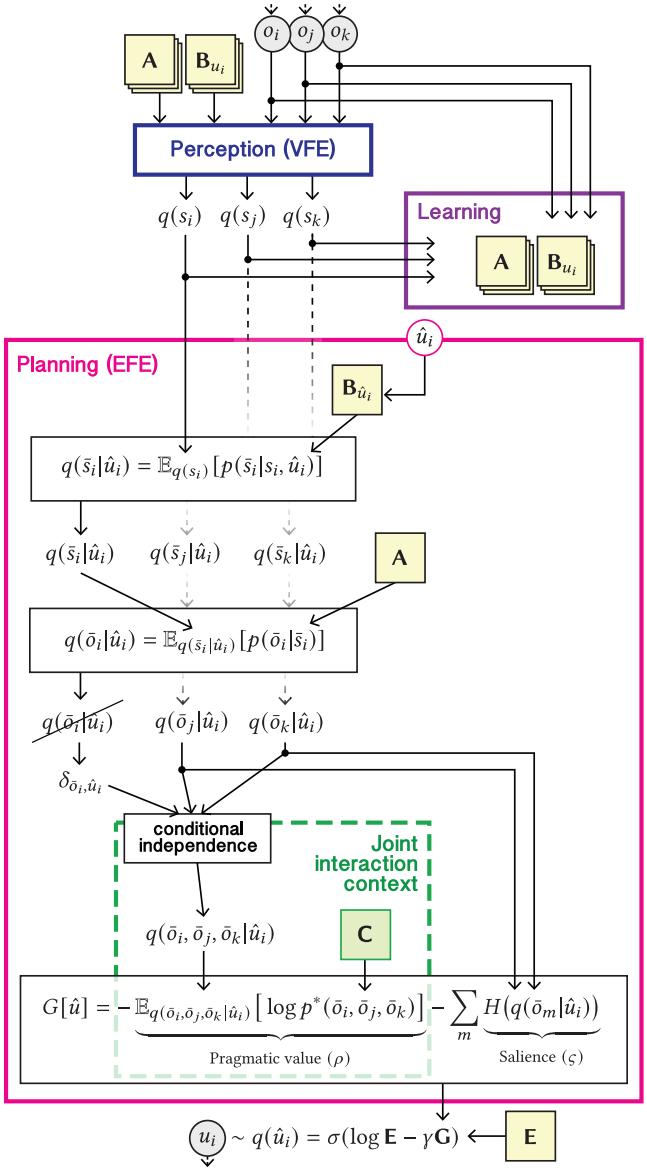

Figure 1: The perception-action loop; ego's 'internal world', ψ_i. The agent observes the actions of all agents, from which she updates her beliefs $q(s)$ to minimise VFE. These beliefs are used to plan her next action by minimising EFE.

3.3.1 Salience Otherwise known as 'epistemic value', salience (ς) captures the information gain about hidden states—or how an action is anticipated to change one's beliefs—with greater changes in beliefs holding higher epistemic value [77]. It can be decomposed into a difference between two entropic terms:

$$\varsigma[\hat{u}] = \mathbb{E}_{q(\bar{o}|\hat{u})}\left[\mathrm{D}_{KL}\left[q(\bar{s}|\hat{u},\bar{o}) \,\middle\|\, q(\bar{s}|\hat{u})\right]\right] \tag{6}$$

$$= H\big(q(\bar{o}|\hat{u})\big) - \mathbb{E}_{q(\bar{s}|\hat{u})}\left[H\big(p(\bar{o}|\bar{s})\big)\right] \tag{7}$$

Since our INFG environment is unambiguous (cf. §3.1), the second term (ambiguity) is zero. There is, however, information to be gained still from acting to maximise the first term, leading to exploratory behaviour. This term captures the uncertainty about the next observation in the event that $\hat{u}_i$ is the action taken. Actions whose outcomes we are most uncertain about are preferred, as we stand to gain the most information from them. Salience is additive over factors:

$$\varsigma[\hat{u}_i] = \sum_m H\big(q(\bar{o}_m|\hat{u}_i)\big) \tag{8}$$

3.3.2 Pragmatic value By choosing actions that maximise pragmatic value (ρ), agents actively pursue their preferences, closing the gap between predicted and preferred observations. In normal-form games, preferences are determined by the game payoffs g, and we can convert directly from one to the other with

$$p^*(o_i, o_j, o_k) = \sigma\big(\mathrm{g}(o_i, o_j, o_k)\big) \tag{9}$$

$\forall (o_i, o_j, o_k) \in \mathcal{U}^N$, where σ is the *softmax* function[5] [17, 73].

This implies a *joint interaction context*, since the preferences for each observation modality come from a joint distribution and are generally not independent. This highlights the core principle of game theory, namely the interdependence of agents' preferences and actions. Here the preferences for each observation modality are derived from a joint distribution, meaning that agents' outcomes are interconnected, and their strategies cannot be considered in isolation. The essence of game theory lies in analyzing how these dependencies shape decision-making and the resulting equilibria in strategic interactions. Agents track the 'mental states' of other agents *individually* (in $q(s_j)$), but they must learn how they interplay in a given joint context defined by g (e.g. when *two* agents play a *Prisoner's Dilemma*, or when *three* agents play *Chicken*). Unlike in classical game theory, however, our agents do not know the payoff function (preferences) of their opponents.

The predicted observations, as a posterior predictive distribution $q(\bar{o}|\hat{u})$ for each modality, need to be merged into a single $q(\bar{o}_i, \bar{o}_j, \bar{o}_k|\hat{u}_i)$ to be able to compare to the preferences. The conditional independence between the internal states of the agents, determined by their respective MBs (§3.2.1), allows us to define this joint posterior as the product [91]

$$q(\bar{o}_i, \bar{o}_j, \bar{o}_k|\hat{u}_i) = \prod_{m\in\{i,j,k\}} q(\bar{o}_m|\hat{u}_i) \tag{10}$$

where each factor's posterior predictive observation is obtained from the (factor's) likelihood model $\mathbf{A}_j$ as[6]

$$q(\bar{o}_j|\hat{u}_i) = \mathbb{E}_{q(\bar{s}_j|\hat{u})}\left[p(\bar{o}_j|\bar{s}_j)\right]. \tag{11}$$

The pragmatic value is thus defined as the negative cross-entropy between the posterior predictive observation and the preference distributions, which agents aim to maximise (cf. log-loss minimisation):

$$\rho[\hat{u}_i] = \mathbb{E}_{q(\bar{o}_i,\bar{o}_j,\bar{o}_k|\hat{u}_i)}\left[\log p^*(\bar{o}_i, \bar{o}_j, \bar{o}_k)\right]. \tag{12}$$

Furthermore, for the ego's own factor, $\bar{o}_i = \hat{u}_i$ is guaranteed with full certainty in the counterfactual where $\hat{u}_i$ is played (i.e. where u_i

[5]The *softmax* operation here is not strictly necessary; we could just as well go with $p^*(o_i, o_j, o_k) = \exp\big(\mathrm{g}(o_i, o_j, o_k)\big)$ under an energy function interpretation. However, by ensuring p^* is a probability distribution, we ensure the EFE values are in the positive range (i.e. in Nats) [37, 59].

[6]$q(\bar{o}_j|\hat{u}_i) = \sum_{\bar{s}_j} q(\bar{o}_j, \bar{s}_j|\hat{u}_i) = \sum_{\bar{s}_j} p(\bar{o}_j|\bar{s}_j) q(\bar{s}_j|\hat{u}_i) = \mathbb{E}_{q(\bar{s}_j|\hat{u})}\left[p(o_j|\bar{s}_j)\right]$

would in actuality be $\hat{u}_i$). Accordingly, we can set $q(\bar{o}_i|\hat{u}_i) = \delta_{\bar{o}_i,\hat{u}_i}$, the Kronecker delta distribution[7]. This makes the pragmatic value

$$\rho[\hat{u}] \equiv \mathbb{E}_{q(\bar{o}_j)q(\bar{o}_k)}[\log p^*(\bar{o}_i, \bar{o}_j, \bar{o}_k)|\bar{o}_i = \hat{u}_i], \tag{13}$$

i.e., equivalent to the game-theoretic *expected utility*, under the interpretation that $\log p^*$ is (proportional to) the utility function (i.e., the game payoffs, as we did in Eq. 9). Finally, we have the EFE,

$$G[\hat{u}_i] = -\rho[\hat{u}_i] - \varsigma[\hat{u}_i], \tag{14}$$

which the agents minimise through their actions, effectively maximising salience (ς) and pragmatic value (ρ). For the zero-ambiguity case under consideration, the EFE reduces to (in shorthand)

$$G[\hat{u}_i] = -\mathbb{E}_{q_{-i}}[\log p^*|\hat{u}_i] - \sum_m H(q_m), \tag{15}$$

highlighting the relationship with previous information-theoretic treatments of bounded rationality in game theory [75, 91].

3.4 Action selection

The selection of actions is driven by the precision-modulated EFE of each possible action $G[\hat{u}_i]$ (or $\mathbf{G}$ in vector notation), and the agent's habits $\mathbf{E}$ (uniform):

$$u_i \sim q(\hat{u}_i) = \sigma\big(\log \mathbf{E} - \gamma \mathbf{G}\big), \tag{16}$$

where γ is a precision parameter updated each time step,

$$\gamma = \frac{\beta_1}{\beta_0 - \langle \mathbf{G} \rangle}, \tag{17}$$

with fixed hyperparameters (β_0, β_1). β_0 (shape) represents a baseline level of uncertainty or noise, and β_1 (rate) reflects how strongly the agent's precision (or confidence) is influenced by environmental feedback. Intuitively, β_0 controls the threshold at which the agent starts doubting its action selections, while β_1 modulates the rate of precision updating based on the difference between expected and observed outcomes. Higher values of β_1 correspond to greater sensitivity to discrepancies, making the agent more 'rational' and noise-averse in refining its action policy [23, 26]. External feedback is accounted for in $\langle \mathbf{G} \rangle = \mathbb{E}_{q(\hat{u}_i)}\big[G[\hat{u}_i]\big]$, the expected EFE under the current action probabilities for this agent.

3.5 Learning

Learning in AIF entails updating model parameters, and occurs at a slower rate than inference. In our model, agents update the transition model $\mathbf{B}$ (initially uniform) every T_L steps, where T_L is an integer sampled uniformly from $18 \leq T_L \leq 30$ each time learning occurs. This random offset ensures agents do not learn in lockstep, which might cause artifacts in the dynamics. The parameters are updated based on past transitions, $h_{t:t+T_L} = (\mathbf{s}_t, u_{i,t}, \mathbf{s}_{t+1}, u_{i,t+1}, ..., \mathbf{s}_{t+T_L})$, via

$$\mathbf{B}'_{u,n} = \mathbf{B}_{u,n} + \sum_{\tau=t}^{t+T_L-1} \alpha_l \, \delta_{u,u_{i,\tau}} (\mathbf{s}_{n,\tau+1} \otimes \mathbf{s}_{n,\tau}) \tag{18}$$

for each factor n, with learning rate $\alpha_l = 1$, and where $\otimes$ is the outer product and $\mathbf{s}_{n,t}$ sufficient statistics for $q(s_n)$ at time t. We refer the reader to [15] for further details.

[7] or, equivalently, $\mathbf{1}(\hat{u}_i)$, the one-hot encoding of the action under consideration

3.5.1 Novelty With learning present, the agents can consider how their actions are likely to influence their generative model. In this case, an agent predicts how the transition model might change if they were to play action $\hat{u}_i$ (for each factor n):

$$\bar{\mathbf{B}}_{\hat{u}_i,n} = \mathbf{B}_{\hat{u}_i,n} + \alpha_l \, (\bar{\mathbf{s}}_n \otimes \mathbf{s}_n) \tag{19}$$

Novelty is an additional term in the EFE, constituting additional epistemic value:

$$\eta[\hat{u}_i] = \sum_n \mathrm{D}_{KL}\left[\bar{\mathbf{B}}_{\hat{u}_i,n} \,\middle\|\, \mathbf{B}_{\hat{u}_i,n}\right] \tag{20}$$

summed over factors. Including novelty in the EFE, we have

$$G[\hat{u}_i] = -\rho[\hat{u}_i] - \varsigma[\hat{u}_i] - \eta[\hat{u}_i] \tag{21}$$

3.5.2 Bayesian model reduction A Bayesian model reduction procedure [15] applied to $\mathbf{B}$ (for each factor and action) at learning time helps reduce overfitting. A 'reduced' model $\tilde{\mathbf{B}}_{u,n} = \sigma(\alpha_r^{-1}\mathbf{B}_{u,n})$ is proposed (with reduction rate $\alpha_r = 1.25$) and their evidence difference is computed as

$$\log \tilde{p}(o_n) - \log p(o_n) = \log \mathbb{E}_{\mathbf{B}'_{u,n}}\left[\frac{\tilde{\mathbf{B}}_{u,n}}{\mathbf{B}_{u,n}}\right]. \tag{22}$$

If the difference is positive (respectively negative), the evidence for the reduced (resp. original) model is greater, so it (resp. the posterior model) is selected as the updated model.

4 Results and discussion

Game transitions increase the non-stationarity of the environment beyond that caused by reciprocal adaptation, resulting in role reversals, strategic uncertainty, and eventual equilibrium selection. We use a time-based linear interpolation of the payoff matrices of two games g and g' to transition between them. Given a time of transition t_x and a transition duration T_x, the preferences p^* of each agent are updated each time step within the interval $(t_\mathrm{x} - \frac{T_\mathrm{x}}{2}) \leq t \leq (t_\mathrm{x} + \frac{T_\mathrm{x}}{2})$, via mixing parameter $l = \big(t - (t_\mathrm{x} - \frac{T_\mathrm{x}}{2})\big)/T_\mathrm{x}$, with

$$p^* = \sigma\Big((1-l)\,\mathsf{g} \;+\; l\,\mathsf{g}'\Big). \tag{23}$$

4.1 VFE and strategic uncertainty

To illustrate the dynamics of transitioning between games, we run a two-player Ch game for 500 iterations, followed by a SH game for another 500, with payoffs as per Eq. 2. The transition occurs over $T_\mathrm{x} = 10$ iterations. The outcome is shown as time series plots for various quantities for each of the two agents in Figure 2.

The VFE, $F[q, o]$, quantifies how surprising an observation o is under the current beliefs $q(s)$. As shown in Fig. 2 (first row)—and further highlighted by the stylised bounds in Fig. 3—, the ensemble is initially in a mixed equilibrium. The symmetry is broken at $t \approx 250$, when, following a model parameter update, i's policy moves towards defection, and j appropriately responds by moving towards cooperation (incentivised by the game payoffs). The ensemble remains in the selected dc equilibrium for the remainder of the Ch game.

If the agents' β_1 values were much higher, the VFE would follow the green stylised curve in Fig. 3 much more closely. However, the chosen $\beta_1 = 15$ value causes some 'suboptimal' actions to be sampled from $q(\hat{u})$ occasionally, so some observations deviate

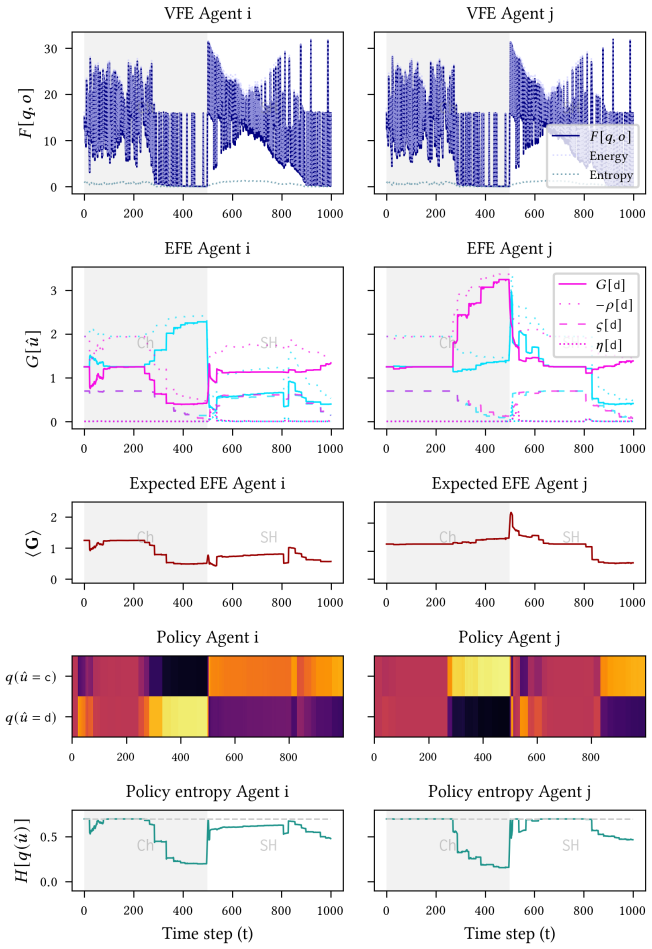

Figure 2: Dynamics of a game transition with two agents ($\beta_1 = 15$). 500 steps of Ch followed by 500 steps of SH with a 10-step transition. In the EFE plots, blue represents 'cooperate', and pink represents 'defect'. In the policy heatmap plots, a lighter colour indicates a higher probability.

from the 'status quo' (Fig. 3, green) in one (orange) or both (red) observation modalities, showing different levels of surprise (VFE) each[8].

After the game transition, there is a role reversal: the cooperating agent becomes a defector, and vice versa. This is explained by the EFE, $G[\hat{u}]$ (Fig. 2, second row), where i's (negative) pragmatic value of cooperating ($-\rho[\mathrm{c}]$, blue) drops dramatically, surpassing the pragmatic value of defecting ($-\rho[\mathrm{d}]$, pink). This is because the preferences p^* have now changed, and because j has been cooperating up to that point—i.e., i believes j is a cooperator, that he can be trusted due to his reputation, making it more appealing for i to cooperate as well. However, the obverse is also at play for j, which transitions to defecting. We note that this role reversal happens regardless of the transition duration, T_{x}. With the transition, there is a sudden increase in the epistemic value (salience ς and novelty η) of actions, leading to exploratory behaviour.

[8]The VFE is additive over the factors of the generative model

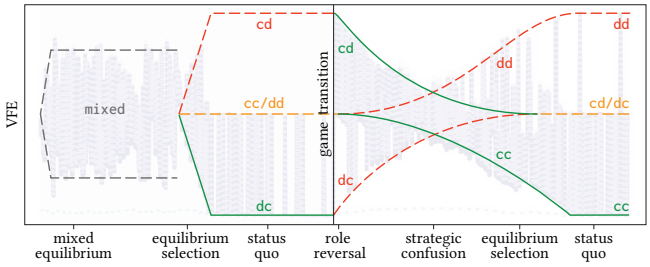

Figure 3: Stylised bounds on the dynamics of the VFE.

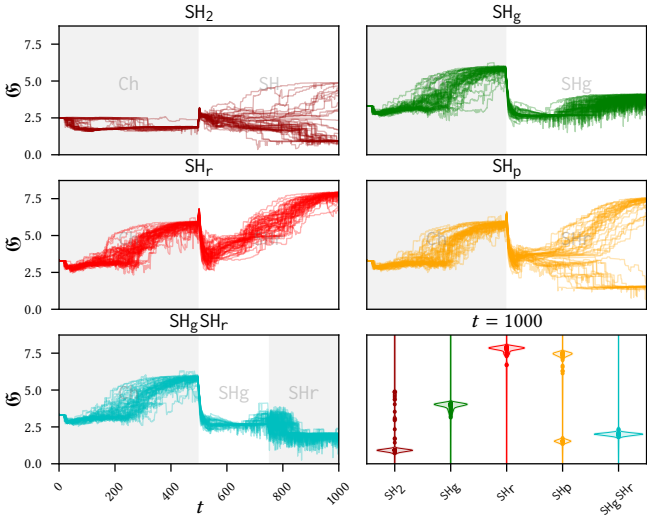

Figure 4: The ensemble-level expected EFE, $\mathfrak{G}$, highlights (the relative size of the basin of attraction of) the equilibria of a game ($\beta_1 = 30$). The bottom-right plot shows the kernel density estimate (Gaussian kernel, 0.08 bandwidth) of the PDF of final values under each condition.

As the agents persist with their new strategies, the pragmatic value of each action changes to reflect them. The absolute difference $|\rho_j[\mathrm{c}] - \rho_j[\mathrm{d}]|$ diminishes, increasing the entropy of the policy, $q(\hat{u}_j)$, up to a point of maximal 'strategic confusion', where the VFE is approximately the same for any possible observation (everything is just as surprising as anything else). This confusion is resolved by equilibrium selection, with a small spike in novelty followed by a decrease in salience as the two agents' policies converge to their final values.

In this particular trial, the ensemble arrives at a payoff-dominant equilibrium with predominantly cc strategies (modulated by the rationality of the agents). But this may not always be the case, as we show next.

4.2 EFE and equilibrium selection

The ensemble-level expected EFE, $\mathfrak{G} = \sum_i \langle \mathbf{G} \rangle^{(i)}$, closely related to the joint objective recently proposed in [47], provides a compact measure of the state of the ensemble over time. By running several trials of an INFG we can use this statistic to characterise the different equilibria of the game, as well as (an approximation of) the relative size of their basin of attraction.

In Figure 4, we show values of $\mathfrak{G}$ under different experimental conditions. The data were obtained by running 50 trials of a chosen (sequence of) INFG for each condition, which in the plots are shown superimposed (for a given condition). Data points that are superimposed on the plots appear darker, highlighting the relative number of trials that take similar values. The chosen INFG are again Ch followed by SH, with a two-player condition (SH_2; as in Fig. 2), and the following three-player conditions using three variants of the SH game:

- a 'green' variant (SH_g) where only two agents are required in order to successfully hunt a stag,
- a 'red' variant (SH_r) where all three agents are required, and
- a 'penalty' variant (SH_p), where all three agents are required and the temptation to defect is lowered (by setting $T = P$).

Their respective payoff matrices are:

$$\mathsf{SH}_\mathsf{g} = \left[\begin{bmatrix} R & R \\ R & S \end{bmatrix}, \begin{bmatrix} T & P \\ P & P \end{bmatrix}\right]; \quad \mathsf{SH}_\mathsf{r} = \left[\begin{bmatrix} R & S \\ S & S \end{bmatrix}, \begin{bmatrix} T & P \\ P & P \end{bmatrix}\right];$$

$$\mathsf{SH}_\mathsf{p} = \left[\begin{bmatrix} R & S \\ S & S \end{bmatrix}, \begin{bmatrix} P & P \\ P & P \end{bmatrix}\right]$$

such that e.g. $g_{\mathsf{SH}_\mathsf{g}}(\mathsf{c},\mathsf{d},\mathsf{c}) = R$, but $g_{\mathsf{SH}_\mathsf{r}}(\mathsf{c},\mathsf{d},\mathsf{c}) = S$, with $R > T > P > S$ assigned the integers from 1 to 4. The payoff matrix for the three-player Ch has the same form as SH_r, except with $T > R > S > P$ (as per §2.1). A fifth and final experimental condition is included, with an additional transition from SH_g to SH_r post-Ch.

The five time-series plots (one for each condition) in Fig. 4 show superimposed series for 50 repeats. The bottom-right plot shows a side-by-side comparison of the values of $\mathfrak{G}$ at $t = 1000$ under each condition.

We first look at the Ch period, where the ensemble starts at a mixed equilibrium ($t = 0$) and reaches a pure equilibrium ($t \leq 500$). In the two-agent case, the ensemble tends toward one of cd or dc fairly quickly, although in one of the trials it stays in the mixed equilibrium through to the end. There are no underlying configuration changes for Ch between the three-agent cases, so any differences are caused by stochasticity. In the final Ch equilibrium, invariably, one of the agents defects and the rest cooperate (up to symmetry). Since $\mathfrak{G}$ is additive, it decreases for the two-player condition, as only one of two agents has to choose the 'worse' action, compared to the increase under the three-player conditions, where two-thirds of the ensemble select the 'worse' action. This is an instance where the Nash Equilibrium is not socially optimal.

The final Ch equilibrium sets the prior conditions for the subsequent SH game, amounting to a 'pre-equilibrium' [94, p. 3]. The SH game has a *risk-dominant equilibrium* (RDE) with a higher $\mathfrak{G}$ (i.e. worse overall), and a *payoff-dominant equilibrium* (PDE) with a lower $\mathfrak{G}$ (i.e. better overall). In the SH_2 and SH_p conditions, we see a bifurcation occur where a portion of the trials ends in each equilibrium, with the majority of SH_2 (resp. SH_p) ending in the PDE (resp. RDE). This shows the relative size of their basins of attraction (noting that SH_2 may need beyond $t = 1000$ to converge further).

All the SH_g trials end in the PDE; all of SH_r, in the RDE. This is somewhat paradoxical: when two are required to cooperate (SH_g), all three cooperate; conversely, when three are required to cooperate (SH_r), none cooperates. Interestingly, the SH_p variant could be seen as an attempt to direct the ensemble towards a better equilibrium via penalising certain behaviour (cf. mechanism design), with mild success. On the other hand, including an interim transition through SH_g generates trust and thus 'bootstraps' cooperative behaviour, stewarding the collective [5] towards the PDE without needing to resort to penalties.

5 Conclusion

We have proposed a factorisation of the generative model of AIF agents that brings the framework in closer alignment with game theory, particularly for multi-agent interactions. In this factorisation, each agent has explicit, individual-level beliefs about the internal states of others, and uses them to plan strategically in a joint (game-theoretic) context. This allows ego to flexibly track the internal states of others outside the joint interaction context, and incorporate that information as required by the interaction. This would be particularly useful when the agents involved in a given interaction change or if the agent participates in multiple interactions at a given time (cf. network games).

We have applied our proposed model to two- and three-agent INFG and included transitions between games that the agents have to adapt to. We have shown how the VFE and EFE can be used to analyse the dynamics of ensembles of agents interacting strategically [7]—in particular, to highlight the equilibria of games and the relative size of their attractors—, and provided an example where this is used to motivate an intervention leading the ensemble to a better outcome based on trust, rather than punishment. The use of these measures may help in the conceptualization of groups of agents as collective agents with self-organizing dynamics and operational closure [47, 55, 79].

AIF and game theory applied to multi-agent systems offer a rich theoretical and experimental landscape for exploring adaptive behaviours in intelligent agent interactions. This intersection of cognitive science and artificial intelligence not only provides insights into individual decision-making but also paves the way for understanding the collective dynamics that shape social behaviour in complex environments.

Ego's beliefs about her own policy $q(\hat{u}_i)$ reflect a form of introspection: inferring one's internal mental states by observing one's actions [43, 85]. This is contrasted with interoception: the perception of internal bodily sensations, which would provide direct access to internal states like $q(\hat{u}_i)$. By endowing ego with interoceptive access to $q(\hat{u}_i)$, one could bypass the need for further inference about her internal state in future steps of the model. Future work shall explore how learning the observation model could capture the rationality of an opponent; the potential for more complex transition models conditioned on the actions of all agents, or for modelling other hidden variables such as opponent preferences [14, 84]; or the effects of different EFE formulations [13, 40, 64] on game outcomes.

Acknowledgments

JRS is supported by an Australian Government Research Training Program (RTP) Scholarship. We acknowledge the Gadi people of the Eora nation as the traditional custodians of the land on which The University of Sydney now stands, and that their sovereignty was never ceded.

References

[1] Stefano V Albrecht, Jacob W Crandall, and Subramanian Ramamoorthy. 2016. Belief and truth in hypothesised behaviours. *Artificial Intelligence* 235 (2016), 63–94.

[2] Stefano V Albrecht and Subramanian Ramamoorthy. 2015. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems. *arXiv preprint arXiv:1506.01170* (2015).

[3] Stefano V Albrecht and Peter Stone. 2018. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence* 258 (2018), 66–95.

[4] Robert Axelrod and William D Hamilton. 1981. The evolution of cooperation. *Science* 211, 4489 (1981), 1390–1396.

[5] Joseph B Bak-Coleman, Mark Alfano, Wolfram Barfuss, Carl T Bergstrom, Miguel A Centeno, Iain D Couzin, Jonathan F Donges, Mirta Galesic, Andrew S Gersick, Jennifer Jacquet, Albert B Kao, Rachel E Moran, Pawel Romanczuk, Daniel I Rubenstein, Kaia J Tombak, Jay J Van Bavel, and Elke U Weber. 2021. Stewardship of Global Collective Behavior. *Proceedings of the National Academy of Sciences* 118, 27 (July 2021), e2025764118. https://doi.org/10.1073/pnas.2025764118

[6] Chris Baker, Rebecca Saxe, and Joshua Tenenbaum. 2011. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 33.

[7] Wolfram Barfuss. 2022. Dynamical Systems as a Level of Cognitive Analysis of Multi-Agent Learning. *Neural Computing and Applications* 34, 3 (Feb. 2022), 1653–1671. https://doi.org/10.1007/s00521-021-06117-0

[8] José M Bernardo and Adrian FM Smith. 2009. *Bayesian theory*. Vol. 405. John Wiley & Sons.

[9] George W Brown. 1951. Iterative Solution of Games by Fictitious Play. In *Activity Analysis of Production and Allocation*, Tjalling C. Koopmans (Ed.). John Wiley & Sons, New York, 374–376.

[10] Colin F Camerer. 2011. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

[11] David Carmel and Shaul Markovitch. 1999. Exploration strategies for model-based learning in multi-agent systems: Exploration strategies. *Autonomous Agents and Multi-agent systems* 2 (1999), 141–172.

[12] Georgios Chalkiadakis and Craig Boutilier. 2003. Coordination in multiagent reinforcement learning: A Bayesian approach. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*. 709–716.

[13] Théophile Champion, Howard Bowman, Dimitrije Marković, and Marek Grześ. 2024. Reframing the Expected Free Energy: Four Formulations and a Unification. arXiv:2402.14460 [cs]

[14] Alex J Chan and M Schaar. 2021. Scalable Bayesian Inverse Reinforcement Learning. In *ICLR*.

[15] Lancelot Da Costa, Thomas Parr, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl J Friston. 2020. Active Inference on Discrete State-Spaces: A Synthesis. *Journal of Mathematical Psychology* 99 (Dec. 2020), 102447. https://doi.org/10.1016/j.jmp.2020.102447

[16] Eddie Dekel, Drew Fudenberg, and David K. Levine. 2004. Learning to play Bayesian games. *Games and Economic Behavior* 46, 2 (2004), 282–303.

[17] Daphne Demekas, Conor Heins, and Brennan Klein. 2024. An Analytical Model of Active Inference in the Iterated Prisoner's Dilemma. In *Active Inference*, Christopher L. Buckley, Daniela Cialfi, Pablo Lanillos, Maxwell Ramstead, Noor Sajid, Hideaki Shimazaki, Tim Verbelen, and Martijn Wisse (Eds.). Springer Nature Switzerland, Cham, 145–172. https://doi.org/10.1007/978-3-031-47958-8_10

[18] Prashant Doshi, Piotr Gmytrasiewicz, and Edmund Durfee. 2020. Recursively modeling other agents for decision making: A research perspective. *Artificial Intelligence* 279 (2020), 103202.

[19] Michael O'Gordon Duff. 2002. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst.

[20] Matthew Fellows, Anuj Mahajan, Tim GJ Rudner, and Shimon Whiteson. 2019. VIREL: A variational inference framework for reinforcement learning. *Advances in Neural Information Processing Systems* 32 (2019).

[21] Chris Fields and James F. Glazebrook. 2024. Nash Equilibria and Undecidability in Generic Physical Interactions—A Free Energy Perspective. *Games* 15, 5 (Oct. 2024), 30. https://doi.org/10.3390/g15050030

[22] Ismael T. Freire, X. Arsiwalla, J. Puigbò, and P. Verschure. 2019. Modeling Theory of Mind in Multi-Agent Games Using Adaptive Feedback Control. *ArXiv* (2019).

[23] Karl J Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, Giovanni Pezzulo, et al. 2016. Active inference and learning. *Neuroscience & Biobehavioral Reviews* 68 (2016), 862–879.

[24] Karl J Friston, Conor Heins, Tim Verbelen, Lancelot Da Costa, Tommaso Salvatori, Dimitrije Markovic, Alexander Tschantz, Magnus Koudahl, Christopher Buckley, and Thomas Parr. 2024. From Pixels to Planning: Scale-Free Active Inference. https://doi.org/10.48550/arXiv.2407.20292 arXiv:2407.20292 [cs, q-bio]

[25] Karl J Friston, Thomas Parr, Conor Heins, Axel Constant, Daniel Friedman, Takuya Isomura, Chris Fields, Tim Verbelen, Maxwell Ramstead, John Clippinger, and Christopher D. Frith. 2024. Federated Inference and Belief Sharing. *Neuroscience & Biobehavioral Reviews* 156 (Jan. 2024), 105500. https://doi.org/10.1016/j.neubiorev.2023.105500

[26] Karl J Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. 2015. Active Inference and Epistemic Value. *Cognitive Neuroscience* 6, 4 (Oct. 2015), 187–214. https://doi.org/10.1080/17588928.2015.1020053

[27] Drew Fudenberg and David K Levine. 1993. Self-confirming equilibrium. *Econometrica: Journal of the Econometric Society* (1993), 523–545.

[28] Drew Fudenberg and David K Levine. 1995. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control* 19, 5-7 (1995), 1065–1089.

[29] Drew Fudenberg and David K. Levine. 1998. *The theory of learning in games*. Vol. 2. MIT Press.

[30] Drew Fudenberg and David K Levine. 1999. Conditional universal consistency. *Games and Economic Behavior* 29, 1-2 (1999), 104–130.

[31] Drew Fudenberg and Jean Tirole. 1991. Repeated games. In *Game Theory*. MIT Press, Cambridge, Massachusetts, 150–192.

[32] Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, Aviv Tamar, et al. 2015. Bayesian reinforcement learning: a survey. *Foundations and Trends® in Machine Learning* 8, 5-6 (2015), 359–483.

[33] Herbert Gintis. 2006. The Foundations of Behavior: The Beliefs, Preferences, and Constraints Model. *Biological Theory* 1, 2 (June 2006), 123–127. https://doi.org/10.1162/biot.2006.1.2.123

[34] Piotr J Gmytrasiewicz and Prashant Doshi. 2004. Interactive POMDPs: Properties and preliminary results. In *International Conference on Autonomous Agents: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, Vol. 3. 1374–1375.

[35] Piotr J Gmytrasiewicz and Prashant Doshi. 2005. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research* 24 (2005), 49–79.

[36] Piotr J Gmytrasiewicz and Edmund H Durfee. 2000. Rational coordination in multi-agent environments. *Autonomous Agents and Multi-Agent Systems* 3 (2000), 319–350.

[37] Sebastian Gottwald and Daniel A Braun. 2020. The Two Kinds of Free Energy and the Bayesian Revolution. *PLOS Computational Biology* 16, 12 (Dec. 2020), e1008420. https://doi.org/10.1371/journal.pcbi.1008420

[38] Peter D Grünwald and A Philip Dawid. 2004. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *the Annals of Statistics* 32, 4 (2004), 1367–1433.

[39] Arthur Guez, David Silver, and Peter Dayan. 2012. Efficient Bayes-adaptive reinforcement learning using sample-based search. *Advances in Neural Information Processing Systems* 25 (2012).

[40] Danijar Hafner, Pedro A Ortega, Jimmy Ba, Thomas Parr, Karl J Friston, and Nicolas Heess. 2022. Action and Perception as Divergence Minimization. https://doi.org/10.48550/arXiv.2009.01791 arXiv:2009.01791 [cs, math, stat]

[41] James Hannan. 1957. Approximation to Bayes Risk in Repeated Play. *Contributions to the Theory of Games* 3 (1957), 97–139.

[42] Michael S Harré. 2018. Multi-Agent Economics and the Emergence of Critical Markets. arXiv:1809.01332 [nlin, q-fin]

[43] Michael S Harré. 2022. What Can Game Theory Tell Us about an AI 'Theory of Mind'? *Games* 13, 3 (June 2022), 46. https://doi.org/10.3390/g13030046

[44] John C Harsanyi. 1967. Games with incomplete information played by "Bayesian" players, I–III Part I. The basic model. *Management science* 14, 3 (1967), 159–182.

[45] Pablo Hernandez-Leal, Michael Kaisers, Tim Baarslag, and Enrique Munoz De Cote. 2017. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183* (2017).

[46] Trong Nghia Hoang and Kian Hsiang Low. 2013. A general framework for interacting Bayes-optimally with self-interested agents using arbitrary parametric model and model prior. *arXiv preprint arXiv:1304.2024* (2013).

[47] David Hyland, Tomáš Gavenčiak, Lancelot Da Costa, Conor Heins, Vojtech Kovarik, Julian Gutierrez, Michael J. Wooldridge, and Jan Kulveit. 2024. Free-Energy Equilibria: Toward a Theory of Interactions Between Boundedly-Rational Agents. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

[48] Edwin T Jaynes. 2003. *Probability theory: The logic of science*. Cambridge University Press.

[49] Albert Jiang and Kevin Leyton-Brown. 2010. Bayesian action-graph games. *Advances in Neural Information Processing Systems* 23 (2010).

[50] Albert Xin Jiang, Kevin Leyton-Brown, and Navin AR Bhat. 2011. Action-graph games. *Games and Economic Behavior* 71, 1 (2011), 141–173.

[51] James S Jordan. 1991. Bayesian learning in normal form games. *Games and Economic Behavior* 3, 1 (1991), 60–81.

[52] Ehud Kalai and Ehud Lehrer. 1993. Rational learning leads to Nash equilibrium. *Econometrica: Journal of the Econometric Society* (1993), 1019–1045.

[53] Hilbert J Kappen, Vicenç Gómez, and Manfred Opper. 2012. Optimal control as a graphical model inference problem. *Machine learning* 87 (2012), 159–182.

[54] Michael Kearns, Michael L Littman, and Satinder Singh. 2013. Graphical models for game theory. *arXiv preprint arXiv:1301.2281* (2013).

[55] Julian Kiverstein, Michael D Kirchhoff, and Tom Froese. 2022. The Problem of Meaning: The Free Energy Principle and Artificial Agency. *Frontiers in Neurorobotics* 16 (June 2022). https://doi.org/10.3389/fnbot.2022.844773

[56] Daphne Koller and Brian Milch. 2001. Structured models for multi-agent interactions. In *Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge*. 233–248.

[57] Daphne Koller and Brian Milch. 2003. Multi-agent influence diagrams for representing and solving games. *Games and Economic Behavior* 45, 1 (2003), 181–221.

[58] Pierfrancesco La Mura and Yoav Shoham. 2013. Expected utility networks. *arXiv preprint arXiv:1301.6714* (2013).

[59] Yann LeCun, Sumit Chopra, Raia Hadsell, Marc Aurelio Ranzato, and Fu Jie Huang. 2006. A Tutorial on Energy-Based Learning. In *Predicting Structured Data*, G. Bakir, T. Hofman, B. Scholkopt, A. Smola, and B. Taskar (Eds.). MIT Press.

[60] Sergey Levine. 2018. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909* (2018).

[61] Tao Li, Yuhan Zhao, and Quanyan Zhu. 2022. The role of information structures in game-theoretic multi-agent learning. *Annual Reviews in Control* 53 (2022), 296–314.

[62] Marlize Lombard and Peter Gärdenfors. 2023. Causal cognition and theory of mind in evolutionary cognitive archaeology. *Biological Theory* 18, 4 (2023), 234–252.

[63] Richard D McKelvey and Thomas R Palfrey. 1995. Quantal response equilibria for normal form games. *Games and Economic Behavior* 10, 1 (1995), 6–38.

[64] Beren Millidge, Alexander Tschantz, and Christopher L. Buckley. 2021. Whence the Expected Free Energy? *Neural Computation* 33, 2 (Feb. 2021), 447–482. https://doi.org/10.1162/neco_a_01354

[65] Michael Moutoussis, Nelson Trujillo-Barreto, Wael El-Deredy, Raymond Dolan, and Karl J Friston. 2014. A Formal Model of Interpersonal Inference. *Frontiers in Human Neuroscience* 8 (2014).

[66] John H Nachbar. 2005. Beliefs in repeated games. *Econometrica* 73, 2 (2005), 459–480.

[67] Brenda Ng, Kofi Boakye, Carol Meyers, and Andrew Wang. 2012. Bayes-adaptive interactive POMDPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26. 1408–1414.

[68] Yaw Nyarko. 1994. Bayesian learning leads to correlated equilibria in normal form games. *Economic Theory* 4 (1994), 821–841.

[69] Brendan O'Donoghue. 2021. Variational Bayesian reinforcement learning with regret bounds. *Advances in Neural Information Processing Systems* 34 (2021), 28208–28221.

[70] Brendan O'Donoghue, Ian Osband, and Catalin Ionescu. 2020. Making sense of reinforcement learning and probabilistic inference. *arXiv preprint arXiv:2001.00805* (2020).

[71] Daniel A Ortega and Pedro A Braun. 2011. Information, utility and bounded rationality. In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3-6, 2011. Proceedings 4*. Springer, 269–274.

[72] Pedro Ortega and Daniel Lee. 2014. An adversarial interpretation of information-theoretic bounded rationality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28.

[73] Pedro A Ortega and Daniel A Braun. 2009. A Conversion between Utility and Information. https://doi.org/10.48550/arXiv.0911.5106 arXiv:0911.5106 [cs, math]

[74] Pedro A Ortega and Daniel A Braun. 2013. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 469, 2153 (2013), 20120683.

[75] Pedro A Ortega, Daniel A Braun, Justin Dyer, Kee-Eung Kim, and Naftali Tishby. 2015. Information-Theoretic Bounded Rationality. arXiv:1512.06789 [cs, math, stat]

[76] Alessandro Panella and Piotr Gmytrasiewicz. 2017. Interactive POMDPs with finite-state models of other agents. *Autonomous Agents and Multi-Agent Systems* 31 (2017), 861–904.

[77] Thomas Parr, Giovanni Pezzulo, and Karl J Friston. 2022. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. The MIT Press. https://doi.org/10.7551/mitpress/12441.001.0001

[78] Judea Pearl. 2014. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier.

[79] Maxwell JD Ramstead, Michael D Kirchhoff, Axel Constant, and Karl J Friston. 2021. Multiscale Integration: Beyond Internalism and Externalism. *Synthese* 198, 1 (Jan. 2021), 41–70. https://doi.org/10.1007/s11229-019-02115-x

[80] Iead Rezek, David S Leslie, Steven Reece, Stephen J Roberts, Alex Rogers, Rajdeep K Dash, and Nicholas R Jennings. 2008. On Similarities between Inference in Game Theory and Machine Learning. *Journal of Artificial Intelligence Research* 33 (Oct. 2008), 259–283. https://doi.org/10.1613/jair.2523

[81] Marc Rigter, Bruno Lacerda, and Nick Hawes. 2021. Risk-averse Bayes-adaptive reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 1142–1154.

[82] Julia Robinson. 1951. An iterative method of solving a game. *Annals of Mathematics* 54, 2 (1951), 296–301.

[83] Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. 2007. Bayes-adaptive POMDPs. *Advances in Neural Information Processing Systems* 20 (2007).

[84] Jaime Ruiz-Serra and Michael S Harré. 2023. Inverse Reinforcement Learning as the Algorithmic Basis for Theory of Mind: Current Methods and Open Problems. *Algorithms* 16, 2 (Feb. 2023), 68. https://doi.org/10.3390/a16020068

[85] Lars Sandved-Smith, Casper Hesp, Jérémie Mattout, Karl J Friston, Antoine Lutz, and Maxwell J D Ramstead. 2021. Towards a Computational Phenomenology of Mental Action: Modelling Meta-Awareness and Attentional Control with Deep Parametric Active Inference. *Neuroscience of Consciousness* 2021, 1 (Jan. 2021), niab018. https://doi.org/10.1093/nc/niab018

[86] Leonard J. Savage. 1954. *The foundations of statistics*. Wiley, New York.

[87] Jeff S Shamma and Gürdal Arslan. 2005. Dynamic fictitious play, dynamic gradient play, and distributed convergence to Nash equilibria. *IEEE Trans. Automat. Control* 50, 3 (2005), 312–327.

[88] Yoav Shoham, Rob Powers, and Trond Grenager. 2007. If Multi-Agent Learning Is the Answer, What Is the Question? *Artificial Intelligence* 171, 7 (May 2007), 365–377. https://doi.org/10.1016/j.artint.2006.02.006

[89] Ryan Smith, Karl J Friston, and Christopher J Whyte. 2022. A Step-by-Step Tutorial on Active Inference and Its Application to Empirical Data. *Journal of Mathematical Psychology* 107 (April 2022), 102632. https://doi.org/10.1016/j.jmp.2021.102632

[90] Finnegan Southey, Michael P Bowling, Bryce Larson, Carmelo Piccione, Neil Burch, Darse Billings, and Chris Rayner. 2012. Bayes' bluff: Opponent modelling in poker. *arXiv preprint arXiv:1207.1411* (2012).

[91] David H Wolpert. 2006. Information Theory — The Bridge Connecting Bounded Rational Game Theory and Statistical Physics. In *Complex Engineered Systems: Science Meets Technology*, Dan Braha, Ali A. Minai, and Yaneer Bar-Yam (Eds.). Springer, Berlin, Heidelberg, 262–290. https://doi.org/10.1007/3-540-32834-3_12

[92] Sarah A Wu, Rose E Wang, James A Evans, Joshua B Tenenbaum, David C Parkes, and Max Kleiman-Weiner. 2021. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science* 13, 2 (2021), 414–432.

[93] Wako Yoshida, Ray J Dolan, and Karl J Friston. 2008. Game Theory of Mind. *PLOS Computational Biology* 4, 12 (Dec. 2008), e1000254. https://doi.org/10.1371/journal.pcbi.1000254

[94] H Peyton Young. 2004. The Interactive Learning Problem. In *Strategic Learning and Its Limits*, H Peyton Young (Ed.). Oxford University Press, 0. https://doi.org/10.1093/acprof:oso/9780199269181.003.0001

[95] H Peyton Young. 2004. *Strategic learning and its limits*. Oxford University Press.